## REMARKS

### I. Preliminary Remarks

Claims 1-54 were pending in the application. Claims 6-28 and 30-54 were withdrawn from consideration as directed to non-elected inventions.

Claims 1-5 and 29 have been amended. Claims 6-28 have been canceled without prejudice to its presentation in future, related applications.

The Applicants request that if the product claims 1-5 and 29 are found novel and non-obvious under 35 U.S.C. § 103(a), methods of using the product (claims 31-54) should be rejoined. *See* 1184 OG 86, (1996).

Upon entry of this amendment claims 1-5 and 29 will be pending.

No new matter has been added.

### II. Priority

In paragraph 5 of the Action, the Examiner alleges that the present claims are not supported in the manner required by 35 U.S.C. § 101 and 112, first paragraph, by the priority application and, therefore, that the present claims are not entitled to the benefit of the filing date of the priority application. The Examiner alleges that the priority application fails to provide any specific, substantial and credible utility and provides no guidance or working examples to teach how to use the claimed invention. The Applicants respectfully disagree.

The crux of the Examiner's rejection of the priority claim is similar to the rejections set forth in the present application, in that the pending claims allegedly lack utility and are not enabled. However, as discussed below, the pending claims have utility and enable a person of skill in the art to make and/or use the claimed invention. Since the prior application's disclosure (U.S. Application Serial No. 09/377,563) is similar to the present application, when the pending claims are found to have utility and be enabled, the prior application must also satisfy the requirements under 35 U.S.C. § 101 and 112, first paragraph. Therefore, the Applicants respectfully request that the effective filing date of the present application be recognized as the filing date of the priority application, August 19, 1999.

**III. The Rejection under 35 U.S.C. § 101 for Lack of Utility should be Withdrawn.**

Claims 1-5 and 29 stand rejected under 35 U.S.C. § 101 because the claimed invention is allegedly not supported by a specific, substantial and credible asserted utility or a well established utility. The Examiner also alleges that the asserted utilities lack "substantial and specific utility because further research to identify or reasonably confirm a "real world" context of use is required." (Office Action, page 3). The Applicants respectfully disagree.

The Utility Examination Guidelines (the "Guidelines") require that a claimed invention have a specific, substantial and credible asserted utility, or, alternatively a well-established utility. The claimed polypeptides are related to the family of olfactory receptors and the fact that the claimed polypeptides share 87% sequence homology with known murine G7 olfactory receptor supports the assignment of the same specific, substantial, and credible utility shared by olfactory receptors to the claimed polypeptides.

Under the Guidelines, Office personnel are instructed to review the specification and claims of the application to determine if a specific and substantial utility that is credible is present. The Guidelines note that the specific and substantial requirement "excludes 'throw-away', insubstantial,' or 'nonspecific' utilities, such as the use of a complex invention as landfill." The Guidelines go on to note that an Examiner's "prima facie showing must establish that it is more likely than not that a person of ordinary skill in the art would not consider that any utility asserted by the Applicants would be specific and substantial." "If the Applicants have asserted that the claimed invention is useful for any particular practical purpose (*i.e.*, it has a 'specific and substantial utility') and the assertion would be considered credible by a person of ordinary skill in the art, do not impose a rejection based on lack of utility." (Guidelines, emphasis added).

The Guidelines also comment on the use of computer based analysis of nucleic acids to assign functions to a nucleic acid or polypeptide based upon homology to sequences found in databases. Specifically, the Guidelines state that the:

> suggestions to adopt a *per se* rule rejecting homology based assertions of utility ***are not adopted***. An applicant is entitled to a patent to the subject matter claimed unless statutory requirements are not met (35 U.S.C. 101, 102, 103, 112) . . .The inquiries involved in assessing utility are fact dependent, and the determinations must be made on the basis of scientific evidence. Reliance on the commenters' *per se* rule, rather than a fact dependent inquiry, is impermissible because the commenters provide no scientific evidence that homology-based assertions of

utility are inherently unbelievable or involve implausible scientific principles. *See, e.g., In re Brana,* 51 F.3d 1560, 1566, 34 USPQ2d 1436, 1441 (Fed. Cir. 1995) (rejection of claims improper where claims did 'not suggest an inherently unbelievable undertaking or involve implausible scientific principles' and where ''prior art * * * discloses structurally similar compounds to those claimed by the applicants which have been proven * * * to be effective').

A patent examiner ***must*** accept a utility asserted by an applicant unless the Office has evidence or sound scientific reasoning to rebut the assertion. The examiner's decision must be supported by a preponderance of all the evidence of record. *In re Oetiker,* 977 F.2d 1443, 1445, 24 USPQ2d 1443,1444 (Fed. Cir. 1992). More specifically, when a patent application claiming a nucleic acid asserts a specific, substantial, and credible utility, and bases the assertion upon homology to existing nucleic acids or proteins having an accepted utility, the asserted utility must be accepted by the examiner unless the Office has sufficient evidence or sound scientific reasoning to rebut such an assertion. ''[A] 'rigorous correlation' need not be shown in order to establish practical utility; 'reasonable correlation' is sufficient.'' *Fujikawa* v. *Wattanasin,* 93 F.3d 1559, 1565, 39 USPQ2d 1895, 1900 (Fed. Cir. 1996). The Office will take into account both the nature and degree of the homology.

When a class of proteins is defined such that the members share a specific, substantial, and credible utility, the reasonable assignment of a new protein to the class of sufficiently conserved proteins would impute the same specific, substantial, and credible utility to the assigned protein. If the preponderance of the evidence of record, or of sound scientific reasoning, casts doubt upon such an asserted utility, the examiner should reject the claim for lack of utility under 35 U.S.C. 101. For example, where a class of proteins is defined by common structural features, but evidence shows that the members of the class do not share a specific, substantial functional attribute or utility, despite having structural features in common, membership in the class may not impute a specific, substantial, and credible utility to a new member of the class. When there is a reason to doubt the functional protein assignment, the utility examination may turn to whether or not the asserted protein encoded by a claimed nucleic acid has a well-established use. If there is a well-established utility for the protein and the claimed nucleic acid, the claim would meet the requirements for utility under 35 U.S.C. 101. If not, the burden shifts to the applicant to provide evidence supporting a well-established utility. There is no *per se* rule regarding homology, and each application must be judged on its own merits.

(Guidelines; emphasis added).

As the Applicants have asserted utilities that are specific, substantial and credible, and well established, the Utility Requirement has been satisfied. The Applicants therefore respectfully request the withdrawal of the rejection under 35 U.S.C. § 101.

**A. The Claimed Invention Has A Specific Utility**

To meet the utility requirement, the invention must be "practically useful," *Anderson v Natta*, 480 F.2d 1392, 1397 (CCPA 1973) and confer a "specific benefit" on the

public. *Brenner v. Manson*, 383 U.S. 519, 534 (1966). The threshold of utility under this standard is not high, and requires merely an "identifiable" benefit. Juicy *Whip Inc. v. Orange Bang Inc.*, 51 USPQ2d 1700 (Fed. Cir. 1999). In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180 (Fed. Cir. 1991), the CAFC explained that "An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility." *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

Inventions that achieve a practical use, a use that is also achieved by other inventions, satisfy the utility requirement. Thus practical utilities can be directed to classes of inventions, so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. *Montedison*, 664 F.2d at 374-75. For example, many materials conduct electricity. This general utility applies to a broad class of inventions (conductive materials) and satisfies the utility requirement of section 101. The fact that other materials also conduct electricity does not mean that other materials that conduct electricity want for utility. What is important, however, is that G protein-coupled receptors (GPCRs) are known to have practical uses well beyond throwaway uses like snake food.

Practical uses for GPCRs include therapeutic and diagnostic uses as well as research-based uses. Many medically significant biological processes are mediated by signal transduction pathways involving G-proteins and other second messengers, and GPCRs are recognized as important therapeutic targets for a wide range of diseases. According to a recently issued United States patent, nearly 350 therapeutic agents targeting GPCRs have been successfully introduced onto the market in only the last fifteen years. (See U.S. Patent No. 6,114,127, at col. 2, lines 45-50.) A recent journal review reported that most GPCR ligands are small and can be mimicked or blocked with synthetic analogues. That, together with the knowledge that numerous GPCRs are targets of important drugs in use today, make identification of GPCRs "a task of prime importance." (See, Marchese *et al.*, *Trends Pharmacol. Sci.*, 20(9): 370-5, 1999). Thus, the allegation that there is no well-established utility for proteins of the class that the Applicants are now claiming is directly refuted by industry evidence.

The Office appears to be under the impression that inventions that are, *inter alia,* useful for use in research, are unpatentable. This is not true. The Patent Office's patent database is replete with patents claiming useful research tools, e.g., spectrophotometers. A material whose only use is as a tool in research may indeed be patentable. *Brenner* excludes only those research purposes where the only use of the material itself is as the subject of research. If *Brenner* had held otherwise, any chemical material would, by virtue of its existence, be useful. However, nowhere do those cases state or imply that a material cannot be patentable if has some other beneficial use in research.

Assay methods, like many other tools used in research, have an immediately realizable "real world" value. For example, an assay method that can identify chemical compounds that possess a particular physical, structural or biological property clearly has "real world" value irrespective and independent from the utility that may be associated with the compounds identified using the assay method. As a consequence, a presumption that assay methods cannot possess utility if the compound isolated or identified using the assay do not have utility would be the product of a flawed analysis of *Brenner*. Such a conclusion also would suggest that processes and products can never possess utility if their utility lies in the field of research. Indeed, the application of this concept of the utility requirement as it relates to methods for assaying or identifying compounds, if taken literally, would mean that claims to methods such as NMR, infrared, x-ray crystallography, and screening for other important biological properties, would be unpatentable because further research would be necessary to establish utility for the compounds identified or assayed. This certainly cannot be the result intended by the Patent Office when issuing the Utility Examination Guidelines.

Genes encoding GPCRs can also be used, for example, for toxicology testing to generate information useful in activities such as drug development, even in cases where little is known as to how a particular GPCR works. No additional experimentation would be required, therefore, to determine whether a GPCR has a practical use as all GPCRs have at least one practical use.

Because all GPCRs, as a class, convey practical benefit (much like the class of DNA ligases identified in the Training Materials), there should be no need to provide additional information about them. A person of ordinary skill in the art need not guess whether any given GPCR conveys a practical benefit. Nor is it necessary to know how or

why any given GPCR works. It is settled law that how or why any invention works is irrelevant to determining utility under 35 U.S.C. §101: "[I]t is not a requirement of patentability that an inventor correctly set forth, or even know, how or why the invention works." *In re Cortright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999)(quoting *Newman v. Quigg*, 877 F.2d 1575, 1581 (Fed. Cir. 1989).

Further, as discussed *infra.* and *supra.*, and as acknowledged by the Office, the claimed polypeptides share at least 87% sequence homology with the murine G7 olfactory receptor. Olfactory receptors, receptors to which the claimed polypeptides have been shown to share sequence homology, are the mechanism in which mammals identify odors. These receptors are known to activate the cAMP cascade, involving G protein α subunits, adenylyl cyclase type III and cyclic nucleotide-gated channels, to transmit odorant signals within the olfactory neuron. (See Schild and Restrepo, *Physiol. Rev.* 78: 429-466, 1998). Once the olfactory neurons are activated, these neurons are depolarized by cation influx through cyclic nucleotide-gated channels that are activated by cAMP and $Ca^{2+}$-activated $Cl^-$ channel current. (Frings *et al.*, *Prog. Neurobiology* 60: 247-289, 2000)

As indicated by the Guidelines, the assignment of the claimed polypeptide to the olfactory receptor family imputes the same specific, substantial, and credible utility to the claimed polypeptide. The Examiner has failed to provide any evidence, less still a preponderance of the evidence, to cast doubt upon any of the asserted utilities.

The Applicants need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner*, 383 U.S. at 532. The amount of evidence required to prove utility depends on the facts of each particular case. *In re Jolles*, 628 F.2d 1322, 1326 (CCPA 1980). "The character and amount of evidence may vary, depending on whether the alleged utility appears to accord with or to contravene established scientific principles and beliefs." Id. Unless there is proof of "total incapacity," or there is a "complete absence of data" to support the Applicants' assertion of utility, the utility requirement is met. *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 (Fed. Cir. 1992); *Envirotec*h, 730 F.2d at 762. The Examiner has failed to provide proof of "total incapacity", and the Applicants have provided information that supports the asserted utilities.

The Examiner is also reminded that a patent applicant's assertion of utility in the disclosure is presumed to be true and correct. *In re Cortright*, 165 F.3d at 1356; *Brana*, 51

F.3d at 1566. If such an assertion is made, the Patent Office bears the burden to demonstrate that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved. Id. To do so, the PTO must provide evidence or sound scientific reasoning. See *In re Langer*, 503 F.2d 1380, 1391-92 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566.

The Applicants have demonstrated a "substantial likelihood" of utility by showing a "reasonable correlation" between the utility of the known composition and the composition being claimed. *Fujikawa v. Wattanasin*, 93 F.3d 1559, 1565 (Fed. Cir. 1996). The presently claimed GPCR is related to known GPCRs. The Examiner has not provided evidence or sound scientific reasoning that one skilled in the art would doubt the "reasonable correlation" advanced by the Applicants.

The present application recites at, for example, pages 35-47 of the specification that the claimed invention can be used, inter alia, to identify ligands, protein binding partners, and/or modulators. Additionally, the polynucleotides of the present invention can be used to generate antibodies useful to localize the polypeptide of the present invention *in vivo* or *in vitro*. The antibodies can also be used to determine the expression pattern of the gene in various tissues which would enable a person of ordinary skill in the art to better understand the function and role of the gene *in vivo*. Thus, there is no question that the Applicants have asserted at least one specific utility and, in fact, have provided numerous specific utilities for the polynucleotides of the present invention. Accordingly, under *Brana*, the Patent Office must accept the utility asserted by the Applicants.

Additionally, the Office appears to be under the assumption that ***absolute*** certainty is required for a polynucleotide to have a specific utility. The standard applicable in this case is not, however, proof to certainty, but rather proof to reasonable probability. As the Supreme Court stated, applicant need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner v. Manson*, 383 U.S. at 532. Although, there may be numerous inventions that may arise from the present application, this standard does not justify the Office's stance that the present invention lacks a specific utility. Thus, the Applicants have complied with the specific utility requirement.

The claimed invention in *Brenner* was directed to a method whose ***only*** utility was making a class of steroids. The disclosure in *Brenner* failed to disclose a utility for the products of that method, which in turn led to a § 101 rejection because the products resulting from the method lacked utility. The Applicants admitted that the products produced by the method would not be patentable if they lacked utility. 148 USPQ 696. The Court stated that the method lacked utility as well, holding:

> We find absolutely no warrant for the proposition that although Congress intended that no patent be granted on a chemical compound whose sole "utility" consists of its potential role as an object of use-testing, a different set of rules was meant to apply to the process which yielded the unpatentable product.

148 USPQ 696.

In *Brenner*, the method of making the compounds, which was the only use recited, was inextricably bound up with the compounds themselves and, as a result, the requirement for utility could not be met until a use for the compounds was found. The Court emphasized that the utility of the claimed invention (*i.e.*, the products) would require further research to identify and ascertain, and the compounds produced by the method would be the object of that research.

In contrast, GPCRs related to known GPCRs stand on a very different basis. As discussed, there are a multitude of utilities for the claimed polypeptides, ***including*** their ability to facilitate research.

The Applicants further assert that long held pre-*Brenner* case law standard supports judging the utility of an invention on whether or not the public derives a benefit from the invention, regardless of how slight the benefit. See, for example, *In re Nelson*, 280 F.2d 172, 178-180 (C.C.P.A. 1960) (stating that "however slight the advantage which the public have received from the inventor, it offers a sufficient reason for his compensation") (citing ROBINSON ON PATENTS (1890)); see also *Lowell v. Lewis*, 1 Mason 182 (Fed. Case. No. 8568, 1817) (stating "if it be more or less useful is... of no importance to the public. If it be not extensively useful it will silently sink into contempt and disregard"). Polypeptides of all types are broadly used in the biotechnology industry, playing key roles in drug and disease discovery processes. Indeed, many such polypeptides enable researchers to find the genes associated with physiological functions. The discovery of such functions readily benefits the public. Accordingly, such tools satisfy the pre-*Brenner* case law standard.

**B. The Claimed Invention Has a Substantial Utility**

The Utility Examination Guidelines also require a claimed invention to have a utility that defines a real-world use (a "substantial utility"). The Applicants teach, as described above, that the claimed invention can be used to make antibodies, identify ligands and other binding partners, such as other proteins that interact with the polypeptide (*i.e.*, a G protein). Thus, it is clear that the claimed invention has real-world uses. All the uses described in the present application are real-world uses and, again, stand in stark contrast to the "throw away" uses (*e.g.*, landfill component or snake food) set forth in the utility guidelines. Thus, there is no question that the Applicants have asserted at least one substantial utility and, in fact, have provided numerous substantial utilities. Accordingly, the Applicants have complied with the substantial utility requirement.

**C. The Claimed Invention Has a Credible Utility**

In addition to a specific and substantial utility, the Utility Examination Guidelines require that such utility be credible (a "credible utility"). That is, whether the assertion of utility is believable to a person of ordinary skill in the art based on the totality of evidence and reasoning provided. Clearly, the numerous specific and substantial utilities asserted by the Applicants are credible.

Assertions of credibility are credible unless "(A) the logic underlying the assertion is seriously flawed, or (B) the facts upon which the assertion is based is inconsistent with the logic underlying the assertion." (See, Revised Interim Utility Guidelines Training Materials.) All the utilities described for the polynucleotide and polypeptide are based on sound logic. Furthermore, the utilities for the claimed polynucleotide are not inconsistent with the logic underlying the assertion that the polynucleotide are useful. Polynucleotides are useful to encode and produce polypeptides to generate antibodies, identify ligands or protein partners, evaluate expression patterns, evaluate protein activity, etc. The Examiner has provided no evidence that the logic is seriously flawed or that the facts upon which these assertions are based are inconsistent with the logic underlying the assertions.

In this respect, the G protein coupled receptor family is analogous to the chemical genus that was the subject of *In re Folkers*, 145 USPQ 390 (CCPA 1965) (Compound that belongs to class of compounds, members of which are recognized as useful, is considered useful under §101). The Patent Office does not serve the public by attempting

to substitute a formulaic analysis of § 101 for the established judgment of the biopharmaceutical industry as to what is "useful." If the Patent Office is aware of any well-grounded scientific literature suggesting that GPCR's are not useful, the Applicants request that it be made of record.

**D. Art-Recognized Utility**

The Utility requirement may also be satisfied by an "Art Established Utility" which means that "a person of ordinary skill in the art would immediately appreciate why the invention is useful based on the characteristics of the invention... and the utility is specific, substantial and credible." (M.P.E.P. §2107).

The Applicants point out that commercial products relating to GPCRs for which no confirmed function has been identified are commercially available. GPCRs, ORF clones of GPCRs, and antibodies that bind to GPCRs are commercially available. For example, the Applicants point out that FabGennix Inc. of Shreveport, Louisiana sells an antibody directed to Retinal Anti-GP75. GPCR75 is said to be a GPCR for which a ligand has not yet been identified (see attached product sheet). Invitrogen sells ORF clones of GPCRs including those for which a ligand has not yet been identified (see attached list, especially noting Clone Ids IOH22483, IOH14039, IOH13056, IOH22637, IOH13239, and IOH13516). MD Bio of Taiwan sells GPCR peptides and antibodies against such polypeptides, again where no ligand has yet been identified. That at least three companies make and sell such GPCR products proves that there is a well-established utility for GPCR's as a class, including the presently claimed GPCR polypeptides. Accordingly there could be no better proof of the utilities of the claimed polypeptides- such products are made by a manufacturer (who expects to sell them) for consumers (who expect to buy them). Any argument that there is no art-recognized utility for such polypeptides seems meritless.

The Applicants also note for the record that the Patent Office apparently agrees with the Applicants' reasoning that GPCRs are useful in that the Office has granted and apparently continues to grant patents to G-protein coupled receptors, their encoding polynucleotides and antibodies directed to them ***in which no natural substrate or specific biological significance*** is ascribed to the GPCR. Specifically, the Applicants would like to bring the following US Patents to the Office's attention:

**6,518,414** Maclennan "Molecular Cloning and Expression of G-Protein Coupled Receptors" (Claims an isolated polynucleotide)
**6,511,826** Li et al. "Polynucleotides Encoding Human G-Protein Chemokine Receptor (CCR5) HDGNR10" (Claims an isolated polynucleotide encoding a protein identified as a "chemokine receptor" with no specific chemokine identified)
**6,372,891** Soppet et al. "Human G-Protein Receptor HPRAJ70" (Claims an antibody directed to a G-protein coupled receptor)
**6,361,967** Agarwal et al. "AXOR10, A G-Protein Coupled Receptor" (Claims an isolated polynucleotide)
**6,348,574** Godiska et al. "Seven Transmembrane Receptors" (Claims an antibody directed to a G-protein coupled receptor)
**6,114,139** Hinuma et al. "G-Protein Coupled Receptor Protein and A DNA Encoding the Receptor" (Claims an isolated polynucleotide).
**6,111,076** Fukusumi et al. "Human G-Protein Coupled Receptor (HIBCD07)" (Claims isolated polypeptide)
**6,107,475** Godiska et al. "Seven Transmembrane Receptors" (Claims isolated polynucleotide and methods)
**6,096,868** Halsey et al. "ECR 673: A 7-Transmembrane G-Protein Coupled Receptor" (Claims isolated polypeptide)
**6,090,575** Li et al. "Polynucleotides Encoding Human G-Protein Coupled Receptor GPR1" (Claims isolated polynucleotide)
**6,071,722** Elshourbagy et al. "Nucleic Acids Encoding A G-Protein Coupled 7TM Receptor (AXOR-1)" (Claims an isolated polynucleotide)
**6,071,719** Halsey et al. "DNA Encoding ECR 673: A 7-Transmembrane G-Protein Coupled Receptor" (Claims an isolated polynucleotide)
**6,060,272** Li et al. "Human G-Protein Coupled Receptors" (Claims isolated polynucleotide)
**6,048,711** Hinuma et al. "Human G-Protein Coupled Receptor Polynucleotides" (Claims isolated polynucleotide)
**6,030,804** Soppet et al. "Polynucleotides Encoding G-Protein Parathyroid Hormone Receptor HLTDG74 Polypeptides" (Claims isolated polynucleotide)
**6,025,154** Li et al. "Polynucleotides Encoding Human G-Protein Chemokine Receptor HDGNR10" (Claims an isolated polynucleotide encoding a protein identified as a "chemokine receptor" with no specific chemokine identified)
**5,998,164** Li et al. "Polynucleotides Encoding Human G-Protein Coupled Receptor GPRZ" (Claims isolated polynucleotide)
**5,994,097** Lal et al. "Polynucleotide Encoding Human G-Protein Coupled Receptor" (Claims isolated polynucleotide)
**5,958,729** Soppet et al. "Human G-Protein Receptor HCEGH45" (Claims isolated polypeptide)
**5,955,309** Ellis et al. "Polynucleotide Encoding G-Protein Coupled Receptor (H7TBA62)" (Claims isolated polynucleotide)
**5,948,890** Soppet et al. "Human G-Protein Receptor HPRAJ70" (Claims isolated polypeptide)
**5,945,307** Glucksmann et al. "Isolated Nucleic Acid Molecules Encoding A G-Protein Coupled Receptor Showing Homology to The 5HT Family of Receptors" (Claims isolated polynucleotide)
**5,942,414** Li et al. Polynucleotides Encoding Human G-Protein Coupled Receptor HIBEF51" (Claims isolated polynucleotide)

**5,912,335** Bergsma et al. "G-Protein Coupled Receptor HUVCT36" (Claims isolated polynucleotide)
**5,874,245** Fukusumi et al. "Human G-Protein Coupled Receptors (HIBCD07)" (Claims isolated polynucleotide)
**5,871,967** Shabon et al. "Cloning of A Novel G-Protein Coupled 7TM Receptor" (Claims isolated polynucleotide)
**5,869,632** Soppet et al. "Human G-Protein Receptor HCEGH45" (Claims isolated polynucleotide)
**5,856,443** MacLennan et al. "Molecular Cloning and Expression of G-Protein Coupled Receptors" (Claims isolated polynucleotide)
**5,834,587** Chan et al. "G-Protein Coupled Receptor, HLTEX11" (Claims isolated polypeptide)
**5,776,729** Soppet et al. "Human G-Protein Receptor HGBER32" (Claims isolated polynucleotide)
**5,763,218** Fujii et al. "Nucleic Acid Encoding Novel Human G-Protein Coupled Receptors" (Claims isolated polynucleotide)
**5,756, 309** Soppet et al. "Nucleic Acid Encoding A Human G-Protein Receptor HPRAJ70 and Method of Producing the Receptor" (Claims isolated polynucleotide)
**5,585,476** MacLennan "Molecular Cloning and Expression of G-Protein Coupled Receptors" (Claims isolated polynucleotide)
**5,759,804** Godiska et al. "Isolated Nucleic Acid Encoding Seven Transmembrane Receptors" (Claims isolated polynucleotide and methods)

The Applicants assert that these issued US Patents are evidence of an art recognized utility for G-protein coupled receptors whose natural ligand is unknown. If the Patent Office's position is that issued patents are ***not*** sufficient evidence of art recognition then the Applicants respectfully request that this position be made of record. In the alternative, if the Patent Office wishes to take the position that these issued patents are directed to non-statutory subject matter, then the Applicants respectfully request that this position be made of record.

The Examiner also alleges that protein belonging to the GPCR family, even if they have similar structures, can have different functions and, therefore, the invention is incomplete. However, the Applicants do not determine function based on the structure of the encoded protein. Rather the prediction is based upon the sequence similarity with known polynucleotides or polypeptides encoded thereby. Although different structures can be formed by different amino acid sequences thereby allowing proteins with similar structures to have different functions, proteins that also share sequence similarity in addition to structural similarity are likely to be part of the protein family. It is well known that the probability that two unrelated polypeptides share more than 40% sequence homology over 70 amino acid residues is exceedingly small. Brenner *et al.*, *Proc. Natl. Acad. Sci.* 95:6073-78 (1998) (See,

attached reference). In the present application homology is in excess of 40% over many more than 70 amino acid residues. The probability, therefore, that the polypeptide encoded by the claimed polynucleotides is related to the reference polypeptides is, accordingly, very high.

The Examiner has failed to provide any references that contradict *Brenner*'s basic rule and has failed to provide any "countervailing evidence" required by the Utility Examination Guidelines. Therefore, the Office has failed to meet its burden in providing evidence indicating that the present invention does not have a substantial, credible, and useful invention.

In view of the foregoing, the Applicants respectfully request that the rejection under 35 U.S.C. § 101 be withdrawn.

**IV. The Rejections under 35 U.S.C. § 112, Second Paragraph should be Withdrawn.**

In paragraph 3 of the Action, the Examiner rejected claim 5 under 35 U.S.C. § 112, second paragraph, for allegedly being indefinite for failing to particularly point out and distinctly claim the subject matter of the invention. In particular, claim 5 recites "CON167" by name without structural limitations. In the foregoing amendment, claim 5 now refers to the amino acid sequence of SEQ ID NO: 2.

In light of the foregoing amendment, the Applicants request the rejection of claim 5 under 35 U.S.C. §112, second paragraph be withdrawn.

**V. The Rejection under 35 U.S.C. § 112, First Paragraph should be Withdrawn.**

In paragraph 4 of the Action, the Examiner rejected claim 5 under 35 U.S.C. § 112, first paragraph for failing to meet to the written description requirement. The Examiner asserted that claim 5 encompasses variants and fragments of proteins whose structure is not known or are different than the function of SEQ ID NO: 2 taught in the specification. The Applicants traverse this rejection.

The domains of a seven transmembrane GPCR are taught at page 4, lines 2-10. In addition, one of skill in the will understand that a seven transmembrane GPCR will have 4 extracellular loops and 4 intracellular loops. The location of these domains can be determined using methods standard in the art, and the functions of the listed domains of a GPCR are well known in the art. (See, Alberts *et al.* Molecular Biology of the Cell 3d pg. 734-735 (attached herewith), Garland Publishing, NY 1994; see also literature cited in

Background section of the specification and Example 5.) The specification enables one to identify domains, fragments, and epitopes. In addition, the prior art does teach one skilled in the art how to predict functional domains of a molecule with computer programs such as "tmstreat.all". Also, methods for determining appropriate peptide sequences to generate antibodies, and methods for screening for antibody specificity, are known in the art, as described in Harlow *et al.* (Antibodies, A Laboratory Manual pp. 75-77 Cold Spring Harbor, 1998).

The claims (claims 1-5 and 28) are limited to fragments that comprise ***epitopes specific*** for CON167. One skilled in the art will reasonably understand that an extracellular domain of a GPCR can be used as an epitope. The specification clearly defines the extracellular loops of CON167 as stated above. The term "epitope specific" finds descriptive support throughout the application, including at p. 3, lines 21-23. The specification provides additional guidance by teaching that extracellular epitopes of CON167 are particularly useful for generating and screening for antibodies. (See specification page 6 lines 14-16). Since one of skill in the art can determine the locations of the extracellular loops by viewing the amino acid sequence of CON167 and using standard techniques in the art, it is clear that the Applicants were in possession of epitopes for antibody recognition.

The Examiner cited *Reagents of the University of California v. Eli Lily & Co.*, 119 F.3d 1559, 43 USPQ 2d 1398 (Fed. Circ. 1997) to indicate the need for either recitation of a representative number of polypeptides falling within the scope of the large genus of polypeptide whose function has yet to be identified. In this case, the University of California adequately disclosed and claimed the cDNA sequence of rat insulin and claimed the cDNA for human insulin. The disclosure of the human insulin cDNA was a prophetic example teaching a method for isolation which lacked the actual nucleotide sequence of human insulin. The Appellate Court determined that the mere recitation of the term cDNA is an inadequate written description of that DNA and it is necessary to describe the cDNA's relevant structure or physical characteristics. In the present application, the full length nucleotide and amino acid sequences of CON167 are fully described which is much more than the teaching of a mere method of isolation. The genus of the claims (domains or epitopes or fragments) are defined by the amino acid sequence taught in the specification, and not some undiscovered sequence. Therefore, the holding in *Reagents* is not applicable to the present invention.

In light of the foregoing remarks, the Applicants request the rejection under 35 U.S.C. § 112, first paragraph should be withdrawn.

**VI. The Rejection Under 35 U.S.C. § 102(e) should be Withdrawn.**

In paragraph 6 of the Action, the Examiner rejected claims 1, 3-5 and 29 under 35 U.S.C. § 102(e) in view of the erroneous assertion that the effective filing date of the instantly claimed invention is August 19, 2000, which is the actual filing date of the instant application.

As discussed above, the effective filing date of the present application is that of its priority application, filed August 19, 1999. The Examiner alleges that the present application is not entitled to its priority date because the prior application did not satisfy the requirements of 35 U.S.C. § 101 and 112, first paragraph. However, the prior application does satisfy the requirements under 35 U.S.C. § 101 and 112, first paragraph, for the reasons set forth above, and therefore is entitled to the priority date of August 19, 1999.

Claims 1-3, 5 and 29 stand rejected under 35 U.S.C. § 102(e) as allegedly anticipated by Stryer *et al.*, U.S. Patent Application Publication NO. 2002/0132273. The effective filing date of Stryer *et al.* is June 22, 2000, which is after the effective date of the present application (August 19, 1999). Therefore, the Stryer reference does not qualify as prior art against the present application.

In view of the foregoing remarks, the Applicants request that the rejection under 35 U.S.C. § 102(e) be withdrawn.

## CONCLUSION

In view of the above, each of the presently pending claims 1-5 and 29 are believed to be in immediate condition for allowance and the Applicants request notification of the same.

Dated: December 30, 2003

Respectfully submitted,

By _____
Sharon M. Sintich
Registration No.: 48,484
MARSHALL, GERSTEIN & BORUN LLP
233 S. Wacker Drive, Suite 6300
Sears Tower
Chicago, Illinois 60606-6357
(312) 474-6300
Agent for Applicant

FabGennix Inc.
INTERNATIONAL




about us
place an order
terms & conditions
site map

2940 Youree Drive. Suite E.
Shreveprot, LA 71104, USA
Customer Service : 318 219 1123
318 798 1704
Fax :318 7981849
Email: Info@FabGennix.com

## PRODUCTS

### Retinal 75 kDa Orphan Receptor

**Anti-GPCR75**

**GPCR75-100P, GPCR75-101AP & GPCR75-112AP**

Recently a novel human G-protein coupled receptor gene has been characterized and mapped to chromosome 2p16. This gene codes for a 540 amino acid protein in retinal pigment epithelium (RPE) and cells surrounding retinal arterioles. In contrast, the Northern blot data obtained from mouse sections suggest the expression of transcripts in photoreceptor inner segments and I outer plexiform layer. The transcripts of the GPCR-75 gene (7kb) are also found in abundance in brain sections. So far, no mutations in GPCR-75 protein were identified in patients suffering from Doynei¦s honeycomb retinal dystrophy (DHRD), an inherited retinal degeneration disease that maps to chromosome 2p16 (1).

The GPCR-75 protein is approximately 78 kDa (540 amino acids) protein that is primarily expressed in human retinal pigment epithelium (RPEs). The GPCR-75 sequence analyses suggest the presence of 7 trans-membrane domains, a characteristic feature of GPCR. The protein has putative N-glycosylation sites near the extra cellular N-terminal end of the proteins. The protein has a large 3 intra cellular loop which might be the site for interaction of G-proteins. The short carboxy terminal is intracellular and has putative post-translational modification lipid modification sites.

The Anti-GPCR-75-selective antibodies were generated against conserved sequences near N- and C-termini of the protein that are unique to GPCR-75 protein. The polyclonal antibody strongly labels a 78 kDa protein in RPE cell extracts. Anti-GPCR-75-selective antibody is also available in affinity-purified form for confocal, Western blotting and immunocytochemical analyses. FabGennix Int. Inc. will also conjugate antibodies with fluorescent probes upon request at extra charge. FabGennix Int. Inc. will also provides antibodies against proteins that are involved in retinal degenerative diseases such as various Anti-PDE antibodies, Anti-MERTK, Anti-Phospho-MERTK, EGF-containing fibulin like intracellular protein (EFEMP1), Anti-Myocilin (TIGR), Anti-Bestrophin, Anti-ELVOL4 and a Usher syndrome specific Anti-USH2a antibodies etc. FabGennix Int. Inc employs cyclic peptide methodology for generating antibodies, which results in higher titer and specificity (2). FabGennix Int. Inc., will also provide Western blot positive controls for most of these antibodies in ready-to-use buffer for easy identification of respective proteins. Limited quantities of antigens are also available. Please enquire for their availability before ordering.

| Catalog # | Host Species | Nature | Cross Reactivity | Quantity/Price |
|---|---|---|---|---|
| GPCR75-100P | Rabbit | Polyclonal Antisera | R,M,H | 100 ul for US $ 195 |

| | | | | |
|---|---|---|---|---|
| GPCR75-101AP | Rabbit | Affinity Purified IgG | R, M, H | 100 ug for US $ 225 |
| GPCR75-112AP | Rabbit | Affinity Purified IgG | R, M, H | 100 ug for US $ 225 |
| PC-GPCR75 | N/A | WB Positive Control | Rat | 5 Applications for US $ 75 |
| P-GPCR75 | N/A | Antigenic Peptides | N/A | 250 ug for US $ 65.00 |

R = rat; M = mouse; H = human; C = chicken; monk = monkey ;
* not all variants are labeled equally

**Immunogen**

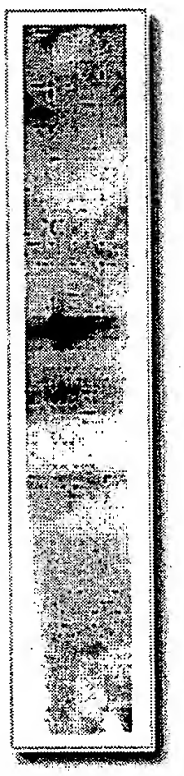
Synthetic cyclic peptide (GPCR75-101AP = PNATSLHVPHSQEGNSTS-amide; GPCR75-112AP = STSLQEGLQDLIHTATLVTC-amide).

**Concentration**

GPCR75-101AP; GPCR-112AP IgG concentration 0.75-1.25 mg/ml in 50% antibody stabilization buffer.

**Applications**

Antibody GPCR75-100/GPCR75-101AP are ideal for WB, IMM and IHC assays. The dilutions for this antibody is for reference only, investigators are expected to determine the optimal conditions for specific assay in his/her laboratory. Dilutions: WB > 1:500; Immunoprecipitation & i.p pull-down assays:> 1:250




Western blot with GPCR75-100P & 65 ug of RPE cell extract. (Ab. dil. 1:500).

**Protocols**

Standard protocol for various applications (WB; IMM and IHC) of this antibody is provided with the product specification sheet, however, FabGennix Int. Inc. strongly recommends investigators to optimize conditions for use of this antibody in their laboratories.

**Form/Storage:**

The antiserum is supplied in antibody stabilization buffer with 0.02% sodium azide or thimerosal/merthiolate as preservative. The affinity-purified antibodies are purified on antigen-speharose affinity column and supplied as 1-1.25 mg/ml IgG in antibody stabilization buffer containing preservatives with low viscosity and cryogenic properties. For long-term storage of antibodies, store at -20oC. Now these antibodies can be stored at iV20oC and used immediately with out thawing. FabGennix Inc. does not recommend storage of very dilute antibody solutions unless they are prepared in specially formulated multi use antibody dilution buffer (Cat # DiluOBuffer). Working solutions of antibodies in DiluOBuffer should be filtered through 0.45ƒÝ filter after every use for long-term storage.

**Discounts**

* For users who may require large amounts of GPCR75-100P or GPCR75-101AP, please enquire about bulk material discounts.

This Product is for Research Use Only and is NOT intended for use in humans or clinical diagnosis.

**Testimonials**

Additional reagents of interest available from FabGennix Inc.

**References:**

1. Tarttelin E. E., Krischner L. S., Bellingham J., Baffi. J. Taymanas S. E., Gre gor E. K., Csaky K., Stratakis C. A., Gregory-Evans C. Y. Biochem. Biophys. Res. Commun. 260, 174-180, 1999.

2. Farooqui, S. M., Brock. W. J., A. Hamdi., Prasad. C. (1991) J. Neurochem. 5 7, 1363-1369.

Back

Back to top ▲

© 2002 Copyright FabGennix Inc. International. All Rights Reserved

Invitrogen™
life technologies

Home | Products & Services | Custom Primers | Technical Resources

Invitrogen Clones

- Invitrogen Clones
- Search Clones
  - CloneRanger™
  - Ultimate™ ORF Browser
  - Important Gene Families
- Clone Collections
  - Ultimate™ ORF Clones
  - PathClones™
  - Genestorm® Clones
  - Full-Length Clones
  - EST Clones
  - BAC/PAC Clones
  - Yeast deletions

home > products & services > invitrogen clones

## Online Catalog - Invitrogen Clones

# Ultimate™ ORF Browser:

## Advanced Search for Ultimate™ ORF Clones

Search By ID or Keyword | Search By Sequence | Browse By Gene Ontology

51 total records for G-Protein Coupled Receptors

| Buy | Clone ID | Species | Definition | Gene Symbol |
|---|---|---|---|---|
| ☐ | IOH4585 | Human | cholecystokinin B receptor | CCKBR |
| ☐ | IOH14244 | Human | Unknown (protein for MGC:23120) | CEACAM1 |
| ☐ | IOH11641 | Human | G-protein coupled receptor (RE2), mRNA. | RE2 |
| ☐ | IOH4151 | Human | pyrimidinergic receptor P2Y6; P2Y6 receptor; G-coupled nucleotide receptor; P2 purinoceptor; P2Y purinoceptor 6 | P2RY6 |
| ☐ | IOH11359 | Human | chemokine (C-X3-C motif) receptor 1; G protein-coupled receptor 13; chemokine (C-C) receptor-like 1; chemokine (C-X3-C) receptor 1 | CX3CR1 |
| ☐ | IOH4363 | Human | guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 2 (GNAT2), mRNA. | GNAT2 |
| ☐ | IOH6739 | Human | melanocortin 1 receptor (alpha melanocyte stimulating hormone receptor) (MC1R), mRNA. | MC1R |
| ☐ | IOH11913 | Human | Similar to urotensin 2 | UTS2 |
| ☐ | IOH13179 | Human | purinergic receptor P2Y, G-protein coupled, 2,transcript variant 1, mRNA (cDNA clone MGC:20088 IMAGE:4561430), | P2RY2 |
| ☐ | IOH14399 | Human | Unknown (protein for MGC:27480) | GPR125 |
| ☐ | IOH3823 | Human | paired box gene 8 isoform PAX8A; paired domain gene 8 | PAX8 |
| ☐ | IOH22614 | Human | G protein-coupled receptor TYMSTR; G protein-coupled receptor | CXCR6 |
| ☐ | IOH22183 | Human | bradykinin receptor B1 (BDKRB1), mRNA. | BDKRB1 |
| ☐ | IOH22483 | Human | clone MGC:33224 IMAGE:5267661, mRNA, complete cds. | RDC1 |

| | | | | |
|---|---|---|---|---|
| ☐ | IOH13929 | Human | dopamine receptor D2 isoform long | DRD2 |
| ☐ | IOH21539 | Human | interleukin 8 receptor, alpha, clone MGC:40015IMAGE:5217529, mRNA, complete cds. | IL8RA |
| ☐ | IOH22790 | Human | G protein-coupled receptor 17, clone MGC:35264IMAGE:5174146, mRNA, complete cds. | GPR17 |
| ☐ | IOH12973 | Human | hypothetical protein MGC24137 (MGC24137), mRNA. | MGC24137 |
| ☐ | IOH22632 | Human | formyl peptide receptor-like 1; lipoxin A4 receptor (formyl peptide receptor related) | FPRL1 |
| ☐ | IOH9895 | Human | G protein-coupled receptor 87 | GPR87 |
| ☐ | IOH3875 | Human | retinoic acid induced 3; retinoic acid responsive gene | RAI3 |
| ☐ | IOH13239 | Human | super conserved receptor expressed in brain 3 | SREB3 |
| ☐ | IOH22669 | Human | adrenomedullin receptor | ADMR |
| ☐ | IOH14201 | Human | endothelial differentiation, sphingolipid G-protein-coupled receptor, 1; edg-1; G protein-coupled sphingolipid receptor; sphingosine 1-phosphate receptor EDG1 | EDG1 |
| ☐ | IOH9916 | Human | coagulation factor II (thrombin) receptor-like 1 | F2RL1 |
| ☐ | IOH1987 | Human | tachykinin receptor 1 isoform short; NK-1 receptor; Tachykinin receptor 1 (substance P receptor; neurokinin-1 receptor); tachykinin 1 receptor (substance P receptor, neurokinin 1 receptor); neurokinin 1 receptor | TACR1 |
| ☐ | IOH22808 | Human | adenosine A3 receptor (ADORA3), mRNA. | ADORA3 |
| ☐ | IOH23130 | Human | similar to olfactory receptor MOR145-1, cloneMGC:32690 IMAGE:4250638, mRNA, complete cds. | LOC115131 |
| ☐ | IOH14039 | Human | Similar to putative nuclear protein ORF1-FL49 | ORF1-FL49 |
| ☐ | IOH9624 | Human | vasoactive intestinal peptide receptor 2 | VIPR2 |
| ☐ | IOH14234 | Human | endothelin receptor type B isoform 1; Hirschsprung disease 2 | EDNRB |
| ☐ | IOH10866 | Human | CD97 antigen isoform 2 precursor; leukocyte antigen CD97; seven-span transmembrane protein | CD97 |
| ☐ | IOH11033 | Human | endothelial differentiation, lysophosphatidic acid G-protein-coupled receptor, 4; G protein-coupled receptor; LPA receptor EDG4; Lysophosphatidic acid receptor EDG4 | EDG4 |
| ☐ | IOH11484 | Human | glycoprotein Ib (platelet), alpha polypeptide | GP1BA |
| ☐ | IOH12342 | Human | olfactory receptor, family 51, subfamily E, member 2 (OR51E2), mRNA. | OR51E2 |
| ☐ | IOH22539 | Human | neuropeptide Y (NPY), mRNA. | NPY |
| ☐ | IOH13127 | Human | EBV-induced G protein-coupled receptor 2; Epstein-Barr virus induced gene 2 | EBI2 |
| ☐ | IOH10344 | Human | Unknown (protein for MGC:21621) | MGC21621 |
| ☐ | IOH10876 | Human | angiotensin receptor 1 | AGTR1 |
| | | Human | | |

| | | | | |
|---|---|---|---|---|
| ☐ | IOH10679 | | endothelin receptor type A | EDNRA |
| ☐ | IOH10700 | Human | somatostatin receptor 2 | SSTR2 |
| ☐ | IOH13056 | Human | similar to POSSIBLE GUSTATORY RECEPTOR CLONE PTE01 | LOC115131 |
| ☐ | IOH3307 | Human | retinal outer segment membrane protein 1 (ROM1), mRNA. | ROM1 |
| ☐ | IOH12543 | Human | purinergic receptor P2Y, G-protein coupled, 12 (P2RY12), transcript variant 1, mRNA. | P2RY12 |
| ☐ | IOH13516 | Human | Similar to G protein-coupled receptor 30 | GPR30 |
| ☐ | IOH10409 | Human | neuromedin U receptor 2 | NMU2R |
| ☐ | IOH3294 | Human | complement component 5 receptor 1 (C5a ligand); complement component-5 receptor-2 (C5a ligand) | C5R1 |
| ☐ | IOH4996 | Human | clone MGC:4303 IMAGE:2819400, mRNA, complete cds. | HTR3A |
| ☐ | IOH13485 | Human | prostaglandin E receptor 3 (subtype EP3), cloneMGC:27302 IMAGE:4660371, mRNA, complete cds. | PTGER3 |
| ☐ | IOH12614 | Human | purinergic receptor P2Y, G-protein coupled, 11 | P2RY11 |
| ☐ | IOH22637 | Human | Similar to parathyroid hormone receptor 1, clone MGC:34562 IMAGE:5180885, mRNA, complete cds. | PTHR1 |

51 total records for G-Protein Coupled Receptors

Add Clones to Shopping Cart

Copyright ©2003 Invitrogen Corporation. All rights reserved. Please read our Privacy Policy. Contact Us.

*Proc. Natl. Acad. Sci. USA*
Vol. 95, pp. 6073–6078, May 1998
Biochemistry

# Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships

STEVEN E. BRENNER*†‡, CYRUS CHOTHIA*, AND TIM J. P. HUBBARD§

*MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, United Kingdom; and §Sanger Centre, Wellcome Trust Genome Campus, Hinxton, Cambs CB10 1SA, United Kingdom

*Communicated by David R. Davies, National Institute of Diabetes, Bethesda, MD, March 16, 1998 (received for review November 12, 1997)*

**ABSTRACT** **Pairwise sequence comparison methods have been assessed using proteins whose relationships are known reliably from their structures and functions, as described in the SCOP database [Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia C. (1995) *J. Mol. Biol.* 247, 536–540]. The evaluation tested the programs BLAST [Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). *J. Mol. Biol.* 215, 403–410], WU-BLAST2 [Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* 266, 460–480], FASTA [Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* 85, 2444–2448], and SSEARCH [Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* 147, 195–197] and their scoring schemes. The error rate of all algorithms is greatly reduced by using statistical scores to evaluate matches rather than percentage identity or raw scores. The E-value statistical scores of SSEARCH and FASTA are reliable: the number of false positives found in our tests agrees well with the scores reported. However, the P-values reported by BLAST and WU-BLAST2 exaggerate significance by orders of magnitude. SSEARCH, FASTA ktup = 1, and WU-BLAST2 perform best, and they are capable of detecting almost all relationships between proteins whose sequence identities are >30%. For more distantly related proteins, they do much less well; only one-half of the relationships between proteins with 20–30% identity are found. Because many homologs have low sequence similarity, most distant relationships cannot be detected by any pairwise comparison method; however, those which are identified may be used with confidence.**

Sequence database searching plays a role in virtually every branch of molecular biology and is crucial for interpreting the sequences issuing forth from genome projects. Given the method's central role, it is surprising that overall and relative capabilities of different procedures are largely unknown. It is difficult to verify algorithms on sample data because this requires large data sets of proteins whose evolutionary relationships are known unambiguously and independently of the methods being evaluated. However, nearly all known homologs have been identified by sequence analysis (the method to be tested). Also, it is generally very difficult to know, in the absence of structural data, whether two proteins that lack clear sequence similarity are unrelated. This has meant that although previous evaluations have helped improve sequence comparison, they have suffered from insufficient, imperfectly characterized, or artificial test data. Assessment also has been problematic because high quality database sequence searching attempts to have both sensitivity (detection of homologs) and specificity (rejection of unrelated proteins); however, these complementary goals are linked such that increasing one causes the other to be reduced.

Sequence comparison methodologies have evolved rapidly, so no previously published tests has evaluated modern versions of programs commonly used. For example, parameters in BLAST (1) have changed, and WU-BLAST2 (2)—which produces gapped alignments—has become available. The latest version of FASTA (3) previously tested was 1.6, but the current release (version 3.0) provides fundamentally different results in the form of statistical scoring.

The previous reports also have left gaps in our knowledge. For example, there has been no published assessment of thresholds for scoring schemes more sophisticated than percentage identity. Thus, the widely discussed statistical scoring measures have never actually been evaluated on large databases of real proteins. Moreover, the different scoring schemes commonly in use have not been compared.

Beyond these issues, there is a more fundamental question: in an absolute sense, how well does pairwise sequence comparison work? That is, what fraction of homologous proteins can be detected using modern database searching methods?

In this work, we attempt to answer these questions and to overcome both of the fundamental difficulties that have hindered assessment of sequence comparison methodologies. First, we use the set of distant evolutionary relationships in the SCOP: Structural Classification of Proteins database (4), which is derived from structural and functional characteristics (5). The SCOP database provides a uniquely reliable set of homologs, which are known independently of sequence comparison. Second, we use an assessment method that jointly measures both sensitivity and specificity. This method allows straightforward comparison of different sequence searching procedures. Further, it can be used to aid interpretation of real database searches and thus provide optimal and reliable results.

**Previous Assessments of Sequence Comparison.** Several previous studies have examined the relative performance of different sequence comparison methods. The most encompassing analyses have been by Pearson (6, 7), who compared the three most commonly used programs. Of these, the Smith–Waterman algorithm (8) implemented in SSEARCH (3) is the oldest and slowest but the most rigorous. Modern heuristics have provided BLAST (1) the speed and convenience to make it the most popular program. Intermediate between these two is FASTA (3), which may be run in two modes offering either greater speed (ktup = 2) or greater effectiveness (ktup = 1). Pearson also considered different parameters for each of these programs.

To test the methods, Pearson selected two representative proteins from each of 67 protein superfamilies defined by the PIR database (9). Each was used as a query to search the database, and the matched proteins were marked as being homologous or unrelated according to their membership of PIR

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

© 1998 by The National Academy of Sciences 0027-8424/98/956073-6$2.00/0
PNAS is available online at http://www.pnas.org.

Abbreviation: EPQ, errors per query.

†Present address: Department of Structural Biology, Stanford University, Fairchild Building D-109, Stanford, CA 94305-5126
‡To whom reprints requests should be addressed. e-mail: brenner@hyper.stanford.edu.

superfamilies. Pearson found that modern matrices and "ln-scaling" of raw scores improve results considerably. He also reported that the rigorous Smith–Waterman algorithm worked slightly better than FASTA, which was in turn more effective than BLAST.

Very large scale analyses of matrices have been performed (10), and Henikoff and Henikoff (11) also evaluated the effectiveness of BLAST and FASTA. Their test with BLAST considered the ability to detect homologs above a predetermined score but had no penalty for methods which also reported large numbers of spurious matches. The Henikoffs searched the SWISS-PROT database (12) and used PROSITE (13) to define homologous families. Their results showed that the BLOSUM62 matrix (14) performed markedly better than the extrapolated PAM-series matrices (15), which previously had been popular.

A crucial aspect of any assessment is the data that are used to test the ability of the program to find homologs. But in Pearson's and the Henikoffs' evaluations of sequence comparison, the correct results were effectively unknown. This is because the superfamilies in PIR and PROSITE are principally created by using the same sequence comparison methods which are being evaluated. Interdependency of data and methods creates a "chicken and egg" problem, and means for example, that new methods would be penalized for correctly identifying homologs missed by older programs. For instance, immunoglobulin variable and constant domains are clearly homologous, but PIR places them in different superfamilies. The problem is widespread: each superfamily in PIR 48.00 with a structural homolog is itself homologous to an average of 1.6 other PIR superfamilies (16).

To surmount these sorts of difficulties, Sander and Schneider (17) used protein structures to evaluate sequence comparison. Rather than comparing different sequence comparison algorithms, their work focused on determining a length-dependent threshold of percentage identity, above which all proteins would be of similar structure. A result of this analysis was the HSSP equation; it states that proteins with 25% identity over 80 residues will have similar structures, whereas shorter alignments require higher identity. (Other studies also have used structures (18–20), but these focused on a small number of model proteins and were principally oriented toward evaluating alignment accuracy rather than homology detection.)

A general solution to the problem of scoring comes from statistical measures (i.e., E-values and P-values) based on the extreme value distribution (21). Extreme value scoring was implemented analytically in the BLAST program using the Karlin and Altschul statistics (22, 23) and empirical approaches have been recently added to FASTA and SSEARCH. In addition to being heralded as a reliable means of recognizing significantly similar proteins (24, 25), the mathematical tractability of statistical scores "is a crucial feature of the BLAST algorithm" (1). The validity of this scoring procedure has been tested analytically and empirically (see ref. 2 and references in ref. 24). However, all large empirical tests used random sequences that may lack the subtle structure found within biological sequences (26, 27) and obviously do not contain any real homologs. Thus, although many researchers have suggested that statistical scores be used to rank matches (24, 25, 28), there have been no large rigorous experiments on biological data to determine the degree to which such rankings are superior.

**A Database for Testing Homology Detection.** Since the discovery that the structures of hemoglobin and myoglobin are very similar though their sequences are not (29), it has been apparent that comparing structures is a more powerful (if less convenient) way to recognize distant evolutionary relationships than comparing sequences. If two proteins show a high degree of similarity in their structural details and function, it is very probable that they have an evolutionary relationship though their sequence similarity may be low.

The recent growth of protein structure information combined with the comprehensive evolutionary classification in the SCOP database (4, 5) have allowed us to overcome previous limitations. With these data, we can evaluate the performance of sequence comparison methods on real protein sequences whose relationships are known confidently. The SCOP database uses structural information to recognize distant homologs, the large majority of which can be determined unambiguously. These superfamilies, such as the globins or the immunoglobulins, would be recognized as related by the vast majority of the biological community despite the lack of high sequence similarity.

From SCOP, we extracted the sequences of domains of proteins in the Protein Data Bank (PDB) (30) and created two databases. One (PDB90D-B) has domains, which were all <90% identical to any other, whereas (PDB40D-B) had those <40% identical. The databases were created by first sorting all protein domains in SCOP by their quality and making a list. The highest quality domain was selected for inclusion in the database and removed from the list. Also removed from the list (and discarded) were all other domains above the threshold level of identity to the selected domain. This process was repeated until the list was empty. The PDB40D-B database contains 1,323 domains, which have 9,044 ordered pairs of distant relationships, or ≈0.5% of the total 1,749,006 ordered pairs. In PDB90D-B, the 2,079 domains have 53,988 relationships, representing 1.2% of all pairs. Low complexity regions of sequence can achieve spurious high scores, so these were masked in both databases by processing with the SEG program (27) using recommended parameters: 12 1.8 2.0. The databases used in this paper are available from http://sss.stanford.edu/sss/, and databases derived from the current version of SCOP may be found at http://scop.mrc-lmb.cam.ac.uk/scop/.

Analyses from both databases were generally consistent, but PDB40D-B focuses on distantly related proteins and reduces the heavy overrepresentation in the PDB of a small number of families (31, 32), whereas PDB90D-B (with more sequences) improves evaluations of statistics. Except where noted otherwise, the distant homolog results here are from PDB40D-B. Although the precise numbers reported here are specific to the structural domain databases used, we expect the trends to be general.

**Assessment Data and Procedure.** Our assessment of sequence comparison may be divided into four different major categories of tests. First, using just a single sequence comparison algorithm at a time, we evaluated the effectiveness of different scoring schemes. Second, we assessed the reliability of scoring procedures, including an evaluation of the validity of statistical scoring. Third, we compared sequence comparison algorithms (using the optimal scoring scheme) to determine their relative performance. Fourth, we examined the distribution of homologs and considered the power of pairwise sequence comparison to recognize them. All of the analyses used the databases of structurally identified homologs and a new assessment criterion.

The analyses tested BLAST (1), version 1.4.9MP, and WU-BLAST2 (2), version 2.0a13MP. Also assessed was the FASTA package, version 3.0t76 (3), which provided FASTA and the SSEARCH implementation of Smith–Waterman (8). For SSEARCH and FASTA, we used BLOSUM45 with gap penalties −12/−1 (7, 16). The default parameters and matrix (BLOSUM62) were used for BLAST and WU-BLAST2.

**The "Coverage Vs. Error" Plot.** To test a particular protocol (comprising a program and scoring scheme), each sequence from the database was used as a query to search the database. This yielded ordered pairs of query and target sequences with associated scores, which were sorted, on the basis of their scores, from best to worst. The ideal method would have

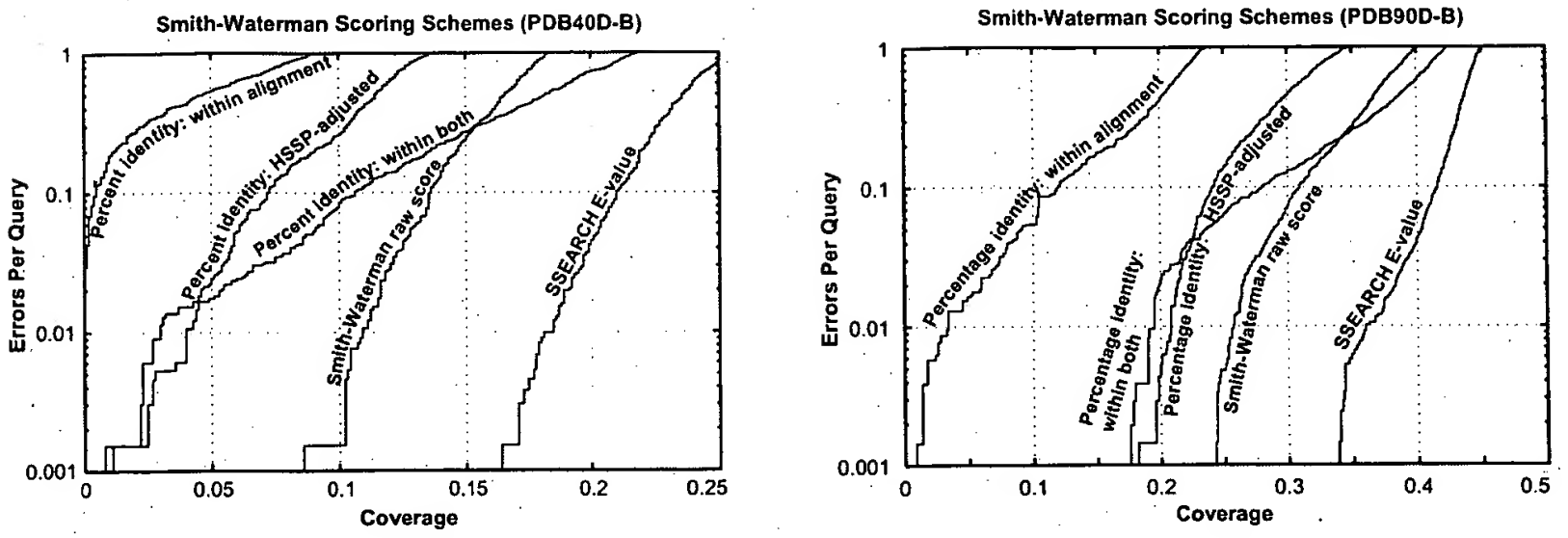



FIG. 1. Coverage vs. error plots of different scoring schemes for SSEARCH Smith–Waterman. (*A*) Analysis of PDB40D-B database. (*B*) Analysis of PDB90D-B database. All of the proteins in the database were compared with each other using the SSEARCH program. The results of this single set of comparisons were considered using five different scoring schemes and assessed. The graphs show the coverage and errors per query (EPQ) for statistical scores, raw scores, and three measures using percentage identity. In the coverage vs. error plot, the $x$ axis indicates the fraction of all homologs in the database (known from structure) which have been detected. Precisely, it is the number of detected pairs of proteins with the same fold divided by the total number of pairs from a common superfamily. PDB40D-B contains a total of 9,044 homologs, so a score of 10% indicates identification of 904 relationships. The $y$ axis reports the number of EPQ. Because there are 1,323 queries made in the PDB40D-B all-vs.-all comparison, 13 errors corresponds to 0.01, or 1% EPQ. The $y$ axis is presented on a log scale to show results over the widely varying degrees of accuracy which may be desired. The scores that correspond to the levels of EPQ and coverage are shown in Fig. 4 and Table 1. The graph demonstrates the trade-off between sensitivity and selectivity. As more homologs are found (moving to the right), more errors are made (moving up). The ideal method would be in the lower right corner of the graph, which corresponds to identifying many evolutionary relationships without selecting unrelated proteins. Three measures of percentage identity are plotted. Percentage identity within alignment is the degree of identity within the aligned region of the proteins, without consideration of the alignment length. Percentage identity within both is the number of identical residues in the aligned region as a percentage of the average length of the query and target proteins. The HSSP equation (17) is $H = 290.15l^{-0.562}$ where $l$ is length for $10 < l < 80$; $H > 100$ for $l < 10$; $H = 24.7$ for $l > 80$. The percentage identity HSSP-adjusted score is the percent identity within the alignment minus H. Smith–Waterman raw scores and E-values were taken directly from the sequence comparison program.

perfect separation, with all of the homologs at the top of the list and unrelated proteins below. In practice, perfect separation is impossible to achieve so instead one is interested in drawing a threshold above which there are the largest number of related pairs of sequences consistent with an acceptable error rate.

Our procedure involved measuring the coverage and error for every threshold. Coverage was defined as the fraction of structurally determined homologs that have scores above the selected threshold; this reflects the sensitivity of a method. Errors per query (EPQ), an indicator of selectivity, is the number of nonhomologous pairs above the threshold divided by the number of queries. Graphs of these data, called coverage vs. error plots, were devised to understand how protocols compare at different levels of accuracy. These graphs share effectively all of the beneficial features of Reciever Operating Characteristic (ROC) plots (33, 34) but better represent the high degrees of accuracy required in sequence comparison and the huge background of nonhomologs.

This assessment procedure is directly relevant to practical sequence database searching, for it provides precisely the information necessary to perform a reliable sequence database search. The EPQ measure places a premium on score consistency; that is, it requires scores to be comparable for different queries. Consistency is an aspect which has been largely

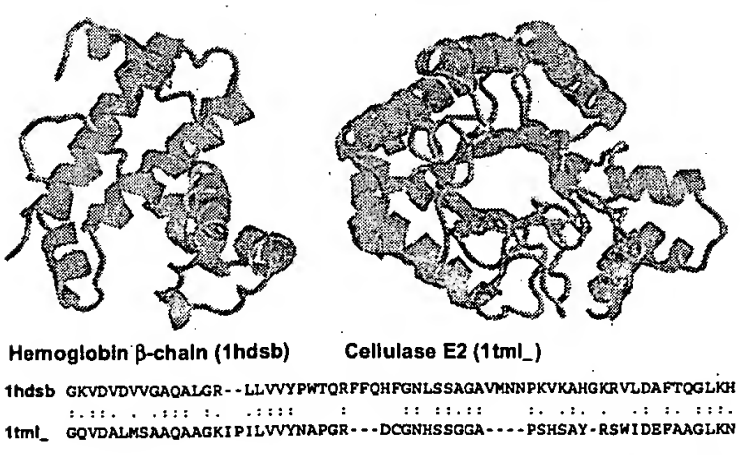



FIG. 2. Unrelated proteins with high percentage identity. Hemoglobin $\beta$-chain (PDB code 1hds chain b, ref. 38, *Left*) and cellulase E2 (PDB code 1tml, ref. 39, *Right*) have 39% identity over 64 residues, a level which is often believed to be indicative of homology. Despite this high degree of identity, their structures strongly suggest that these proteins are not related. Appropriately, neither the raw alignment score of 85 nor the E-value of 1.3 is significant. Proteins rendered by RASMOL (40).

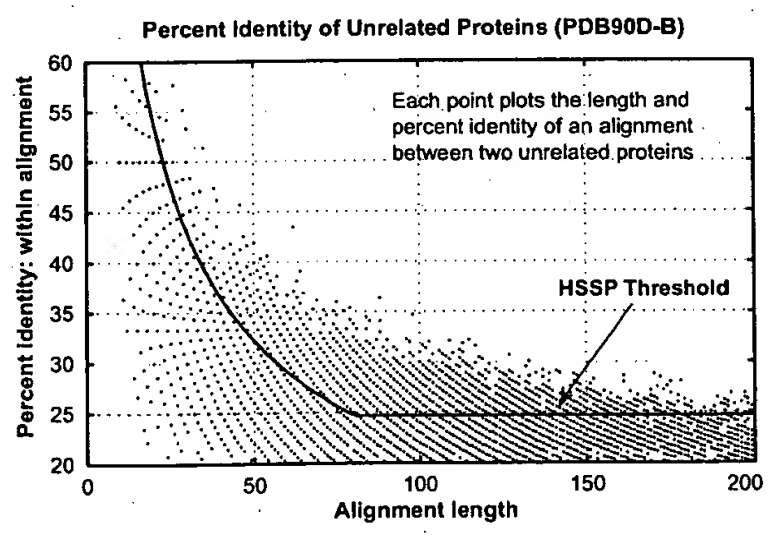



FIG. 3. Length and percentage identity of alignments of unrelated proteins in PDB90D-B: Each pair of nonhomologous proteins found with SSEARCH is plotted as a point whose position indicates the length and the percentage identity within the alignment. Because alignment length and percentage identity are quantized, many pairs of proteins may have exactly the same alignment length and percentage identity. The line shows the HSSP threshold (though it is intended to be applied with a different matrix and parameters).

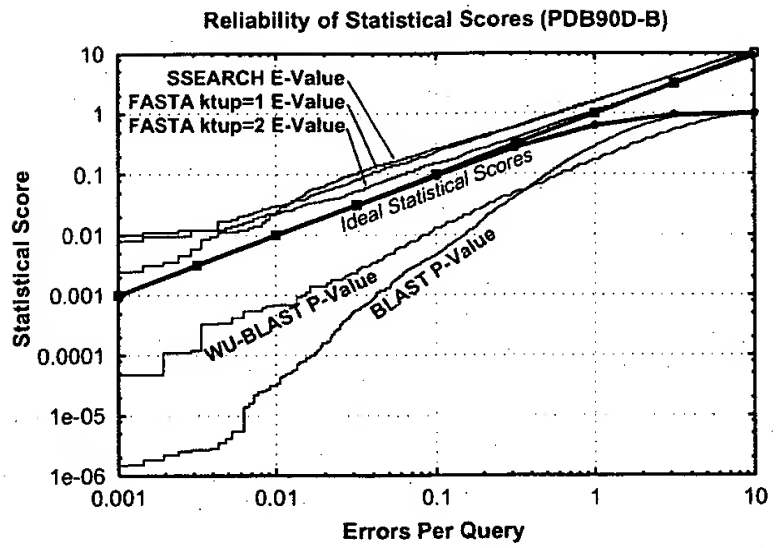



FIG. 4. Reliability of statistical scores in PDB90D-B: Each line shows the relationship between reported statistical score and actual error rate for a different program. E-values are reported for SSEARCH and FASTA, whereas P-values are shown for BLAST and WU-BLAST2. If the scoring were perfect, then the number of errors per query and the E-values would be the same, as indicated by the upper bold line. (P-values should be the same as EPQ for small numbers, and diverges at higher values, as indicated by the lower bold line.) E-values from SSEARCH and FASTA are shown to have good agreement with EPQ but underestimate the significance slightly. BLAST and WU-BLAST2 are overconfident, with the degree of exaggeration dependent upon the score. The results for PDB40D-B were similar to those for PDB90D-B despite the difference in number of homologs detected. This graph could be used to roughly calibrate the reliability of a given statistical score.

ignored in previous tests but is essential for the straightforward or automatic interpretation of sequence comparison results. Further, it provides a clear indication of the confidence that should be ascribed to each match. Indeed, the EPQ measure should approximate the expectation value reported by database searching programs, if the programs' estimates are accurate.

**The Performance of Scoring Schemes.** All of the programs tested could provide three fundamental types of scores. The first score is the percentage identity, which may be computed in several ways based on either the length of the alignment or the lengths of the sequences. The second is a "raw" or "Smith–Waterman" score, which is the measure optimized by the Smith–Waterman algorithm and is computed by summing the substitution matrix scores for each position in the alignment and subtracting gap penalties. In BLAST, a measure related to this score is scaled into bits. Third is a statistical score based on the extreme value distribution. These results are summarized in Fig. 1.

**Sequence Identity.** Though it has been long established that percentage identity is a poor measure (35), there is a common rule-of-thumb stating that 30% identity signifies homology. Moreover, publications have indicated that 25% identity can be used as a threshold (17, 36). We find that these thresholds, originally derived years ago, are not supported by present results. As databases have grown, so have the possibilities for chance alignments with high identity; thus, the reported cutoffs lead to frequent errors. Fig. 2 shows one of the many pairs of proteins with very different structures that nonetheless have high levels of identity over considerable aligned regions. Despite the high identity, the raw and the statistical scores for such incorrect matches are typically not significant. The principal reasons percentage identity does so poorly seem to be that it ignores information about gaps and about the conservative or radical nature of residue substitutions.

From the PDB90D-B analysis in Fig. 3, we learn that 30% identity is a reliable threshold for this database only for sequence alignments of at least 150 residues. Because one unrelated pair of proteins has 43.5% identity over 62 residues, it is probably necessary for alignments to be at least 70 residues in length before 40% is a reasonable threshold, for a database of this particular size and composition.

At a given reliability, scores based on percentage identity detect just a fraction of the distant homologs found by statistical scoring. If one measures the percentage identity in the aligned regions without consideration of alignment length, then a negligible number of distant homologs are detected. Use of the HSSP equation improves the value of percentage identity, but even this measure can find only 4% of all known homologs at 1% EPQ. In short, percentage identity discards most of the information measured in a sequence comparison.

**Raw Scores.** Smith–Waterman raw scores perform better than percentage identity (Fig. 1), but ln-scaling (7) provided no notable benefit in our analysis. It is necessary to be very precise when using either raw or bit scores because a 20% change in cutoff score could yield a tenfold difference in EPQ. However, it is difficult to choose appropriate thresholds because the reliability of a bit score depends on the lengths of the proteins matched and the size of the database. Raw score thresholds also are affected by matrix and gap parameters.

**Statistical Scores.** Statistical scores were introduced partly to overcome the problems that arise from raw scores. This scoring scheme provides the best discrimination between homologous proteins and those which are unrelated. Most

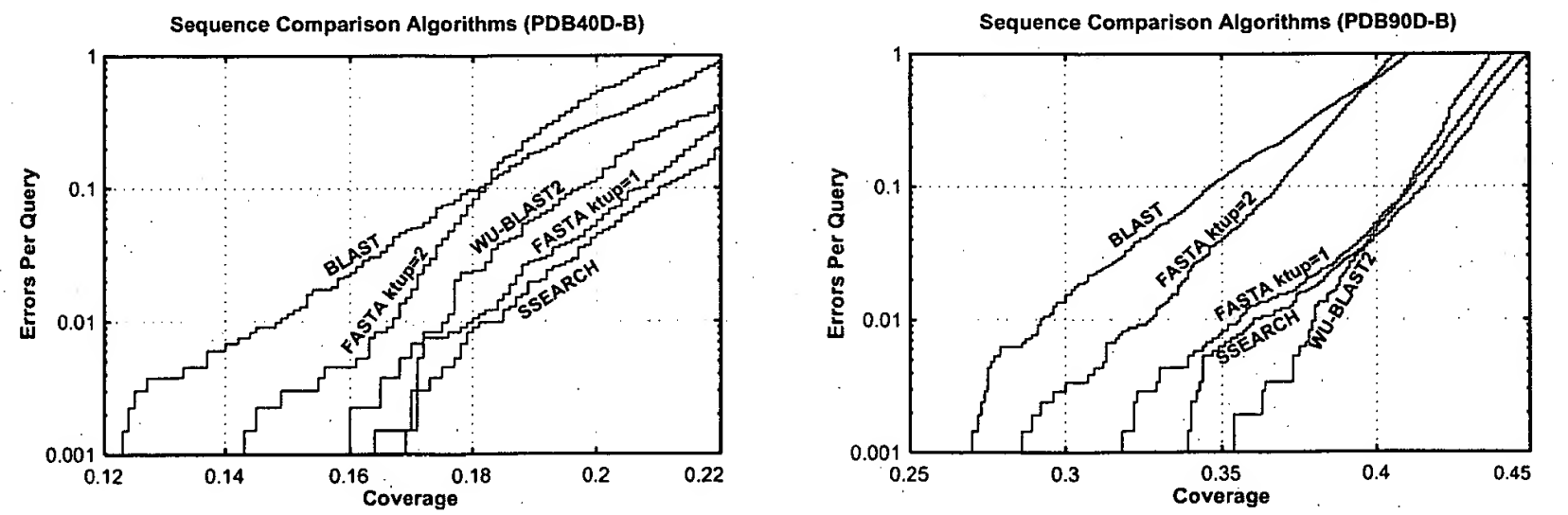



FIG. 5. Coverage vs. error plots of different sequence comparison methods: Five different sequence comparison methods are evaluated, each using statistical scores (E- or P-values). (*A*) PDB40D-B database. In this analysis, the best method is the slow SSEARCH, which finds 18% of relationships at 1% EPQ. FASTA ktup = 1 and WU-BLAST2 are almost as good. (*B*) PDB90D-B database. The quick WU-BLAST2 program provides the best coverage at 1% EPQ on this database, although at higher levels of error it becomes slightly worse than FASTA ktup = 1 and SSEARCH.

likely, its power can be attributed to its incorporation of more information than any other measure; it takes account of the full substitution and gap data (like raw scores) but also has details about the sequence lengths and composition and is scaled appropriately.

We find that statistical scores are not only powerful, but also easy to interpret. SSEARCH and FASTA show close agreement between statistical scores and actual number of errors per query (Fig. 4). The expectation value score gives a good, slightly conservative estimate of the chances of the two sequences being found at random in a given query. Thus, an E-value of 0.01 indicates that roughly one pair of nonhomologs of this similarity should be found in every 100 different queries. Neither raw scores nor percentage identity can be interpreted in this way, and these results validate the suitability of the extreme value distribution for describing the scores from a database search.

The P-values from BLAST also should be directly interpretable but were found to overstate significance by more than two orders of magnitude for 1% EPQ for this database. Nonetheless, these results strongly suggest that the analytic theory is fundamentally appropriate. WU-BLAST2 scores were more reliable than those from BLAST, but also exaggerate expected confidence by more than an order of magnitude at 1% EPQ.

**Overall Detection of Homologs and Comparison of Algorithms.** The results in Fig. 5*A* and Table 1 show that pairwise sequence comparison is capable of identifying only a small fraction of the homologous pairs of sequences in PDB40D-B. Even SSEARCH with E-values, the best protocol tested, could find only 18% of all relationships at a 1% EPQ. BLAST, which identifies 15%, was the worst performer, whereas FASTA ktup = 1 is nearly as effective as SSEARCH. FASTA ktup = 2 and WU-BLAST2 are intermediate in their ability to detect homologs. Comparison of different algorithms indicates that those capable of identifying more homologs are generally slower. SSEARCH is 25 times slower than BLAST and 6.5 times slower than FASTA ktup = 1. WU-BLAST2 is slightly faster than FASTA ktup = 2, but the latter has more interpretable scores.

In PDB90D-B, where there are many close relationships, the best method can identify only 38% of structurally known homologs (Fig. 5*B*). The method which finds that many relationships is WU-BLAST2. Consequently, we infer that the differences between FASTA kup = 1, SSEARCH, and WU-BLAST2 programs are unlikely to be significant when compared with variation in database composition and scoring reliability.

Fig. 6 helps to explain why most distant homologs cannot be found by sequence comparison: a great many such relationships have no more sequence identity than would be expected by chance. SSEARCH with E-values can recognize >90% of the homologous pairs with 30–40% identity. In this region, there are 30 pairs of homologous proteins that do not have significant E-values, but 26 of these involve sequences with <50 residues. Of sequences having 25–30% identity, 75% are identified by SSEARCH E-values. However, although the number of homologs grows at lower levels of identity, the detection falls off sharply: only 40% of homologs with 20–25% identity

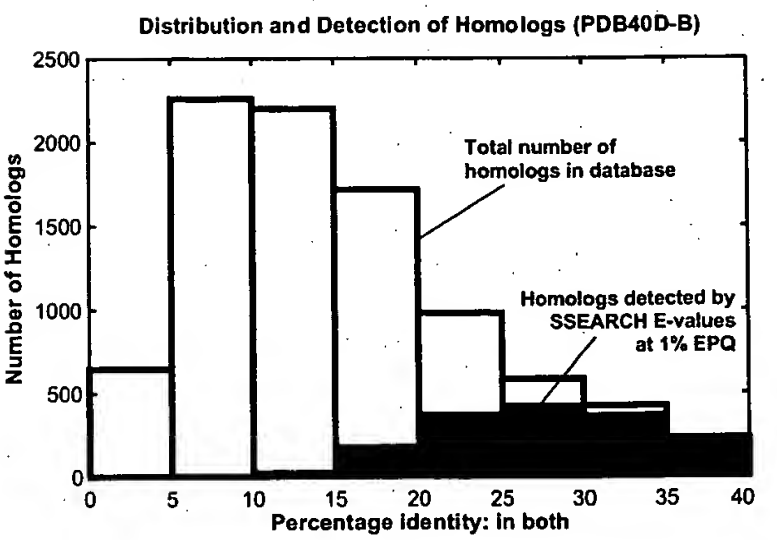



FIG. 6. Distribution and detection of homologs in PDB40D-B. Bars show the distribution of homologous pairs PDB40D-B according to their identity (using the measure of identity in both). Filled regions indicate the number of these pairs found by the best database searching method (SSEARCH with E-values) at 1% EPQ. The PDB40D-B database contains proteins with <40% identity, and as shown on this graph, most structurally identified homologs in the database have diverged extremely far in sequence and have <20% identity. Note that the alignments may be inaccurate, especially at low levels of identity. Filled regions show that SSEARCH can identify most relationships that have 25% or more identity, but its detection wanes sharply below 25%. Consequently, the great sequence divergence of most structurally identified evolutionary relationships effectively defeats the ability of pariwise sequence comparison to detect them.

are detected and only 10% of those with 15–20% can be found. These results show that statistical scores can find related proteins whose identity is remarkably low; however, the power of the method is restricted by the great divergence of many protein sequences.

After completion of this work, a new version of pairwise BLAST was released: BLASTGP (37). It supports gapped alignments, like WU-BLAST2, and dispenses with sum statistics. Our initial tests on BLASTGP using default parameters show that its E-values are reliable and that its overall detection of homologs was substantially better than that of ungapped BLAST, but not quite equal to that of WU-BLAST2.

## CONCLUSION

The general consensus amongst experts (see refs. 7, 24, 25, 27 and references therein) suggests that the most effective sequence searches are made by (*i*) using a large current database in which the protein sequences have been complexity masked and (*ii*) using statistical scores to interpret the results. Our experiments fully support this view.

Our results also suggest two further points. First, the E-values reported by FASTA and SSEARCH give fairly accurate estimates of the significance of each match, but the P-values provided by BLAST and WU-BLAST2 underestimate the true

Table 1. Summary of sequence comparison methods with PDB40D-B

| Method | Relative Time* | 1% EPQ Cutoff | Coverage at 1% EPQ |
|---|---|---|---|
| SSEARCH % identity: within alignment | 25.5 | >70% | <0.1 |
| SSEARCH % identity: within both | 25.5 | 34% | 3.0 |
| SSEARCH % identity: HSSP-scaled | 25.5 | 35% (HSSP + 9.8) | 4.0 |
| SSEARCH Smith–Waterman raw scores | 25.5 | 142 | 10.5 |
| SSEARCH E-values | 25.5 | 0.03 | 18.4 |
| FASTA ktup = 1 E-values | 3.9 | 0.03 | 17.9 |
| FASTA ktup = 2 E-values | 1.4 | 0.03 | 16.7 |
| WU-BLAST2 P-values | 1.1 | 0.003 | 17.5 |
| BLAST P-values | 1.0 | 0.00016 | 14.8 |

*Times are from large database searches with genome proteins.

extent of errors. Second, SSEARCH, WU-BLAST2, and FASTA ktup = 1 perform best, though BLAST and FASTA ktup = 2 detect most of the relationships found by the best procedures and are appropriate for rapid initial searches.

The homologous proteins that are found by sequence comparison can be distinguished with high reliability from the huge number of unrelated pairs. However, even the best database searching procedures tested fail to find the large majority of distant evolutionary relationships at an acceptable error rate. Thus, if the procedures assessed here fail to find a reliable match, it does not imply that the sequence is unique; rather, it indicates that any relatives it might have are distant ones.**

**Additional and updated information about this work, including supplementary figures, may be found at http://sss.stanford.edu/sss/.

The authors are grateful to Drs. A. G. Murzin, M. Levitt, S. R. Eddy, and G. Mitchison for valuable discussion. S.E.B. was principally supported by a St. John's College (Cambridge, UK) Benefactors' Scholarship and by the American Friends of Cambridge University. S.E.B. dedicates his contribution to the memory of Rabbi Albert T. and Clara S. Bilgray.

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990) *J. Mol. Biol.* **215,** 403–410.
2. Altschul, S. F. & Gish, W. (1996) *Methods Enzymol.* **266,** 460–480.
3. Pearson, W. R. & Lipman, D. J. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 2444–2448.
4. Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995) *J. Mol. Biol.* **247,** 536–540.
5. Brenner, S. E., Chothia, C., Hubbard, T. J. P. & Murzin, A. G. (1996) *Methods Enzymol.* **266,** 635–643.
6. Pearson, W. R. (1991) *Genomics* **11,** 635–650.
7. Pearson, W. R. (1995) *Protein Sci.* **4,** 1145–1160.
8. Smith, T. F. & Waterman, M. S. (1981) *J. Mol. Biol.* **147,** 195–197.
9. George, D. G., Hunt, L. T. & Barker, W. C. (1996) *Methods Enzymol.* **266,** 41–59.
10. Vogt, G., Etzold, T. & Argos, P. (1995) *J. Mol. Biol.* **249,** 816–831.
11. Henikoff, S. & Henikoff, J. G. (1993) *Proteins* **17,** 49–61.
12. Bairoch, A. & Apweiler, R. (1996) *Nucleic Acids Res.* **24,** 21–25.
13. Bairoch, A., Bucher, P. & Hofmann, K. (1996) *Nucleic Acids Res.* **24,** 189–196.
14. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
15. Dayhoff, M., Schwartz, R. M. & Orcutt, B. C. (1978) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. (National Biomedical Research Foundation, Silver Spring, MD), Vol. 5, Suppl. 3, pp. 345–352.
16. Brenner, S. E. (1996) Ph.D. thesis. (University of Cambridge, UK).
17. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
18. Johnson, M. S. & Overington, J. P. (1993) *J. Mol. Biol.* **233,** 716–738.
19. Barton, G. J. & Sternberg, M. J. E. (1987) *Protein Eng.* **1,** 89–94.
20. Lesk, A. M., Levitt, M. & Chothia, C. (1986) *Protein Eng.* **1,** 77–78.
21. Arratia, R., Gordon, L. & M, W. (1986) *Ann. Stat.* **14,** 971–993.
22. Karlin, S. & Altschul, S. F. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 2264–2268.
23. Karlin, S. & Altschul, S. F. (1993) *Proc. Natl. Acad. Sci. USA* **90,** 5873–5877.
24. Altschul, S. F., Boguski, M. S., Gish, W. & Wootton, J. C. (1994) *Nat. Genet.* **6,** 119–129.
25. Pearson, W. R. (1996) *Methods Enzymol.* **266,** 227–258.
26. Lipman, D. J., Wilbur, W. J., Smith, T. F. & Waterman, M. S. (1984) *Nucleic Acids Res.* **12,** 215–226.
27. Wootton, J. C. & Federhen, S. (1996) *Methods Enzymol.* **266,** 554–571.
28. Waterman, M. S. & Vingron, M. (1994) *Stat. Science* **9,** 367–381.
29. Perutz, M. F., Kendrew, J. C. & Watson, H. C. (1965) *J. Mol. Biol.* **13,** 669–678.
30. Abola, E. E., Bernstein, F. C., Bryant, S. H., Koetzle, T. F. & Weng, J. (1987) in *Crystallographic Databases: Information Content, Software Systems, Scientific Applications*, eds. Allen, F. H., Bergerhoff, G. & Sievers, R. (Data Comm. Intl. Union Crystallogr., Cambridge, UK), pp. 107–132.
31. Brenner, S. E., Chothia, C. & Hubbard, T. J. P. (1997) *Curr. Opin. Struct. Biol.* **7,** 369–376.
32. Orengo, C., Michie, A., Jones S, Jones D. T, Swindells M. B. & Thornton, J. (1997) *Structure (London)* **5,** 1093–1108.
33. Zweig, M. H. & Campbell, G. (1993) *Clin. Chem.* **39,** 561–577.
34. Gribskov, M. & Robinson, N. L. (1996) *Comput. Chem.* **20,** 25–33.
35. Fitch, W. M. (1966) *J. Mol. Biol.* **16,** 9–16.
36. Chung, S. Y. & Subbiah, S. (1996) *Structure (London)* **4,** 1123–1127.
37. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
38. Girling, R., Schmidt, W., Jr, Houston, T., Amma, E. & Huisman, T. (1979) *J. Mol. Biol.* **131,** 417–433.
39. Spezio, M., Wilson, D. & Karplus, P. (1993) *Biochemistry* **32,** 9906–9916
40. Sayle, R. A. & Milner-White, E. J. (1995) *Trends Biochem. Sci.* **20,** 374–376.

# MOLECULAR BIOLOGY OF
# THE CELL
## THIRD EDITION

Bruce Alberts • Dennis Bray
Julian Lewis • Martin Raff • Keith Roberts
James D. Watson

Garland Publishing, Inc.
New York & London

The complexity of such signal-response systems, with multiple interacting relay chains of signaling proteins, is daunting. But recombinant DNA technology, combined with classical genetic analyses in *Drosophila*, the nematode *C. elegans*, and yeasts, as well as more conventional biochemical and pharmacological methods, is rapidly uncovering the intricate details of these mechanisms by which activated receptor proteins change the behavior of the cell.

## Summary

***Each cell in a multicellular animal is programmed during development to respond to a specific set of signals that act in various combinations to regulate the behavior of the cell and to determine whether the cell lives or dies and whether it proliferates or stays quiescent. Most of these signals mediate paracrine signaling, in which local mediators are rapidly taken up, destroyed, or immobilized, so that they act only on neighboring cells. In addition, centralized control is exerted both by endocrine signaling, in which hormones secreted by endocrine cells are carried in the blood to target cells throughout the body, and by synaptic signaling, in which neurotransmitters secreted by nerve cells act locally on the postsynaptic cells that their axons contact.***

***Cell signaling requires both extracellular signaling molecules and a complementary set of receptor proteins in each cell that enable it to bind and respond to them in a programmed and characteristic way. Some small hydrophobic signaling molecules, including the steroid and thyroid hormones and the retinoids, diffuse across the plasma membrane of the target cell and activate intracellular receptor proteins, which directly regulate the transcription of specific genes. Some dissolved gases, such as nitric oxide and carbon monoxide, act as local mediators by diffusing across the plasma membrane of the target cell and activating an intracellular enzyme—usually guanylyl cyclase, which produces cyclic GMP in the target cell. But most extracellular signaling molecules are hydrophilic and are able to activate receptor proteins only on the surface of the target cell; these receptors act as signal transducers, converting the extracellular binding event into intracellular signals that alter the behavior of the target cell. There are three main families of cell-surface receptors, each of which transduces extracellular signals in a different way. Ion-channel-linked receptors are transmitter-gated ion channels that open or close briefly in response to the binding of a neurotransmitter. G-protein-linked receptors indirectly activate or inactivate plasma-membrane-bound enzymes or ion channels via trimeric GTP-binding proteins (G proteins). Enzyme-linked receptors either act directly as enzymes or are associated with enzymes; the enzymes are usually protein kinases that phosphorylate specific proteins in the target cell. Through cascades of highly regulated protein phosphorylations, elaborate sets of interacting proteins relay most signals from the cell surface to the nucleus, thereby altering the cell's pattern of gene expression and, as a consequence, its behavior. Cross-talk between different signaling cascades enables a cell to integrate information from the multiple signals that it receives.***

## Signaling via G-Protein-linked Cell-Surface Receptors [11]

**G-protein-linked receptors** are the largest family of cell-surface receptors. More than 100 members have already been defined in mammals. Many of these have been identified by *homology cloning*, in which low stringency hybridization with existing cDNA probes is used to detect related DNA sequences (see Figure 7–17). Other family members have been found by *expression cloning*, using their ligand-binding or cell-activation properties to identify them. In one form of this approach, a library of cDNA molecules prepared from cells or tissues that express the receptor are copied into RNA molecules, which are then injected into *Xenopus* oocytes. The oocytes translate the RNA molecules into proteins. These proteins

are inserted into the plasma membrane, where their ligand-binding or cell-activation properties allow them to be detected.

G-protein-linked receptors mediate the cellular responses to an enormous diversity of signaling molecules, including hormones, neurotransmitters, and local mediators, which are as varied in structure as they are in function: the list includes proteins and small peptides, as well as amino acid and fatty acid derivatives. The same ligand can activate many different family members. At least 9 distinct G-protein-linked receptors are activated by adrenaline, for example, another 5 or more by acetylcholine, and at least 15 by serotonin.

Despite the chemical and functional diversity of the signaling molecules that bind to them, all of the G-protein-linked receptors whose amino acid sequences are known from DNA sequencing studies have a similar structure and are almost certainly evolutionarily related. They consist of a single polypeptide chain that threads back and forth across the lipid bilayer seven times (Figure 15–17). As we discuss later, this superfamily of seven-pass transmembrane receptor proteins includes *rhodopsin,* the light-activated protein in the vertebrate eye, as well as olfactory receptors in the vertebrate nose. Other family members are found in unicellular organisms: the receptors in yeasts that recognize the yeast mating factors are an example. This ancient structural motif is also shared by bacteriorhodopsin, a bacterial light-activated $H^+$ pump discussed in Chapter 10, although, unlike the other family members, bacteriorhodopsin is not a receptor and does not act via a G protein. Taken together, these findings suggest that the G-protein-linked receptors that mediate cell-cell signaling in multicellular organisms may have evolved from sensory receptors possessed by their unicellular ancestors. The members of this receptor family have conserved not only their amino acid sequence but also their functional relationship to G proteins by means of which they broadcast into the interior of the cell the message that an extracellular ligand is present. It is the intracellular sequence of events beginning with the activation of G proteins that mainly concern us in this section.

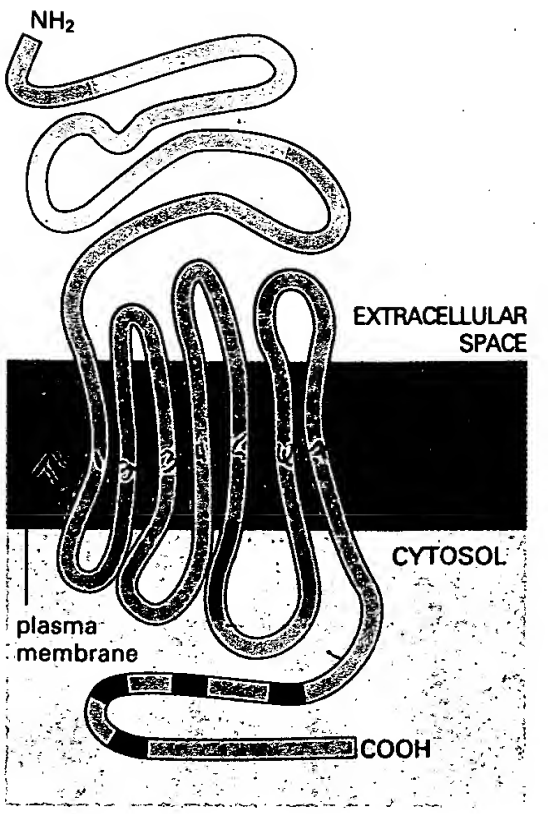



Figure 15–17 **A schematic drawing of a G-protein-linked receptor.** Receptors that bind protein ligands have a large extracellular ligand-binding domain formed by the part of the polypeptide chain shown in *light green.* Receptors for small ligands such as adrenaline have small extracellular domains, and the ligand-binding site is usually deep within the plane of the membrane, formed by amino acids from several of the transmembrane segments. The parts of the intracellular domains that are mainly responsible for binding to trimeric G proteins are shown in *orange,* while those that become phosphorylated during receptor desensitization (discussed later) are shown in *red.*

## Trimeric G Proteins Relay the Intracellular Signal from G-Protein-linked Receptors [11, 12]

The **trimeric GTP-binding proteins (G proteins)** that functionally couple these receptors to their target enzymes or ion channels in the plasma membrane are structurally distinct from the single-chain GTP-binding proteins (called *monomeric GTP-binding proteins* or *monomeric GTPases*) that help relay intracellular signals and regulate vesicular traffic and many other processes in eucaryotic cells. The monomeric GTPases are discussed later in this chapter as well as in other chapters. Both classes of GTP-binding proteins, however, are GTPases and function as molecular switches that can flip between two states: active, when GTP is bound, and inactive, when GDP is bound. "Active" in this context usually means that the molecule acts as a signal to trigger other events in the cell. When an extracellular ligand binds to a G-protein-linked receptor, the receptor changes its conformation and switches on the trimeric G proteins that associate with it by causing them to eject their GDP and replace it with GTP. The switch is turned off when the G protein hydrolyzes its own bound GTP, converting it back to GDP. But before that occurs, the active protein has an opportunity to diffuse away from the receptor and deliver its message for a prolonged period to its downstream target.

Most G-protein-linked receptors activate a chain of events that alters the concentration of one or more small intracellular signaling molecules. These small molecules, often referred to as **intracellular mediators** (also called *intracellular messengers* or *second messengers*), in turn pass the signal on by altering the behavior of selected cellular proteins. Two of the most widely used intracellular mediators are *cyclic AMP (cAMP)* and $Ca^{2+}$: changes in their concentrations are stimulated by distinct pathways in most animal cells, and most G-protein-linked receptors regulate one or the other of them, as outlined in Figure 15–18.